engineering proceedings

MDPI

Proceeding Paper

Uncertainty-Aware IoT Intrusion Detection Using Neutrosophic Ensemble Classification: Disentangling Confidence Magnitude from Uncertainty Geometry

Maikel Leyva-Vázquez [1,2,3,a], Lorenzo Cevallos-Torres [2,b], Alexis Matheu Pérez [3,c] and Florentin Smarandache [4,d]

Published: date

[1] Universidad de Guayaquil, Guayaquil, Ecuador
[2] Universidad Bolivariana del Ecuador, Guayaquil, Ecuador
[3] Universidad Bernardo O'Higgins, Santiago, Chile
[4] University of New Mexico, Gallup, NM, USA
[a] maikel.leyvav@ug.edu.ec, [b] lorenzo.cevallost@ug.edu.ec, [c] alexis.matheu@ubo.cl, [d] smarand@unm.edu
* Correspondence: alexis.matheu@ubo.cl
† Presented at the 2026 IEEE 6th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB 2026), Tamkang University, New Taipei, Taiwan, 24–26 April 2026.

Abstract: Machine learning-based Intrusion Detection Systems (IDS) for IoT environments typically produce scalar confidence scores that fail to characterize the nature of prediction uncertainty. Neutrosophic logic offers a richer framework by decomposing predictions into truth (T), falsity (F), and indeterminacy (I), but a fundamental question remains: does I provide information beyond F? This paper proposes a neutrosophic ensemble framework integrating Random Forest, XGBoost, and Logistic Regression with three independent indeterminacy metrics, evaluated on the IoT-CAD dataset (82,332 instances). We formally prove that I_{max} = F algebraically, eliminating a redundant metric. Through logistic regression ($\Delta AIC = 573$, $p < 10^{-127}$), band-stratified analysis, and matched-pair comparisons, we demonstrate that I_{margin} contributes significant nonlinear incremental value beyond F, concentrated at the critical decision boundary ($F \in [0.3, 0.5]$) where instances with high I_{margin} exhibit 24.4% error versus 8.8% for low I_{margin} ($p < 10^{-114}$). These findings disentangle two independent dimensions: confidence magnitude (T/F) determines whether a prediction is uncertain; uncertainty geometry (I) determines how dangerous that uncertainty is. The ensemble achieves 97% accuracy with selective abstention reaching 99.2%.

Keywords: neutrosophic logic; ensemble classification; indeterminacy; uncertainty quantification; intrusion detection; IoT security; edge computing; trust-aware AI; abstention

1. Introduction

Intrusion Detection Systems (IDS) are critical for IoT security, where heterogeneous devices create expansive attack surfaces [1]. Machine learning-based IDS achieve high accuracy but produce deterministic outputs that fail to characterize prediction uncertainty [2]. An IDS assigning 51% probability to an attack triggers the same automated response as one with 99% confidence—a dangerous conflation in edge deployments where automated responses carry real consequences [3].

Neutrosophic logic, introduced by Smarandache [4], decomposes propositions into truth (T), falsity (F), and indeterminacy (I), distinguishing between confident predictions, predictions against the chosen class, and genuinely ambiguous predictions. Al-Masri [5] applied this decomposition in the NeutroSENSE framework for ensemble-based IDS, demonstrating that misclassified samples exhibit higher indeterminacy.

However, a fundamental question remains unaddressed: does indeterminacy provide predictive information beyond falsity? Since T = max(P) and F = 1 − T, falsity already encodes the complement of confidence. If I merely recapitulates this information, the neutrosophic tripleta reduces to one effective dimension, undermining the three-valued decomposition.

This paper addresses this gap. We propose a neutrosophic ensemble framework for IoT intrusion detection and provide the first systematic analysis disentangling confidence magnitude (T/F) from uncertainty geometry (I). Our contributions are: (1) formal proof that $I_{max} = F$, eliminating a redundant metric; (2) first empirical evidence that I adds nonlinear predictive value beyond F; (3) demonstration that this value concentrates at the critical decision boundary where abstention matters most; (4) discovery that entropy-based and margin-based indeterminacy capture qualitatively opposite uncertainty regimes.

2. Background and Related Work

Neutrosophic logic [4] extends fuzzy logic with three independent membership functions: T, I, F ∈ [0, 1], where their sum is unconstrained. Kavitha et al. [6] applied neutrosophic reasoning to IDS with threshold optimization via genetic algorithms. Kraipeerapun et al. [7] combined interval neutrosophic sets with bagging neural networks. Akbulut et al. [8] introduced NWELM for imbalanced classification. Al-Masri [5] proposed NeutroSENSE with modern ensemble classifiers and adaptive abstention. None addressed whether I provides information beyond F.

Ensemble methods are standard in IoT-IDS [1]. Soft voting preserves distributional information [9], but conventional ensembles collapse it into point predictions [10]. Deep ensembles [10] address uncertainty but are computationally expensive for edge deployment. Neutrosophic decomposition offers a lightweight alternative extracting uncertainty directly from the probability distribution.

In any formulation where T = max(P) and F = 1 − T, the metric I_{max} = 1 − max(P) = F by algebraic identity. We exclude it and retain only metrics capturing distributional geometry: normalized entropy ($I_{entropy}$), margin gap (I_{margin}), and normalized Gini impurity (I_{gini}).

3. Proposed Framework

The framework integrates ensemble classification with neutrosophic uncertainty decomposition. Raw IoT traffic is preprocessed (normalization, encoding, SMOTE balancing), then classified by three complementary models: Random Forest, XGBoost, and Logistic Regression. Soft-voted probability distributions are decomposed into neutrosophic components (T, I, F). An adaptive abstention module flags high-uncertainty predictions for human review.

https://doi.org/10.3390/xxxxx

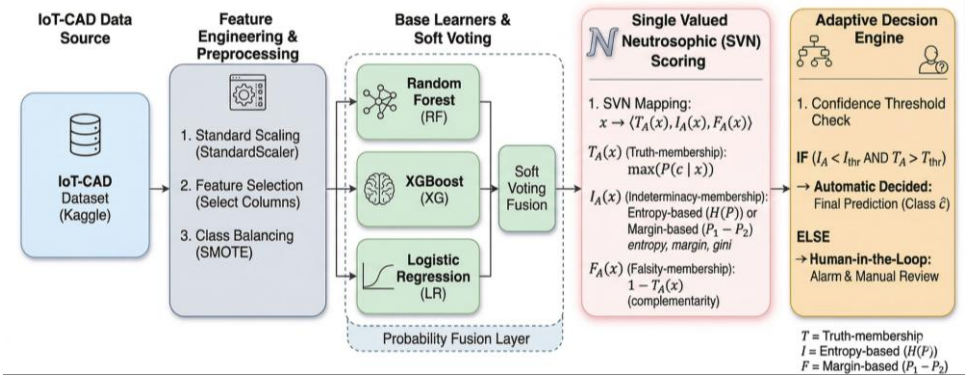
IoT-CAD Data Source
IoT-CAD Dataset (Kaggle)
Feature Engineering & Preprocessing
1. Standard Scaling (StandardScaler)
2. Feature Selection (Select Columns)
3. Class Balancing (SMOTE)
Base Learners & Soft Voting
Random Forest (RF)
XGBoost (XG)
Logistic Regression (LR)
Soft Voting Fusion
Probability Fusion Layer
Single Valued Neutrosophic (SVN) Scoring
1. SVN Mapping:
$x \to \langle T_A(x), I_A(x), F_A(x) \rangle$
$T_A(x)$ (Truth-membership): $\max(P(c \mid x))$
$I_A(x)$ (Indeterminacy-membership): Entropy-based ($H(P)$) or Margin-based ($P_1 - P_2$)
entropy, margin, gini
$F_A(x)$ (Falsity-membership): $1 - T_A(x)$ (complementarity)
Adaptive Decsion Engine
1. Confidence Threshold Check
IF ($I_A < I_{thr}$ AND $T_A > T_{thr}$)
→ Automatic Decided: Final Prediction (Class $\hat{c}$)
ELSE
→ Human-in-the-Loop: Alarm & Manual Review
T = Truth-membership
I = Entropy-based ($H(P)$)
F = Margin-based ($P_1 - P_2$)

Figure 1. Architecture of the proposed neutrosophic ensemble framework for IoT intrusion detection

The framework is implemented in the Orange Data Mining visual programming environment [11], enabling reproducibility through shareable workflows. Figure 2 shows the implemented workflow canvas, where each processing stage corresponds to a visual widget.

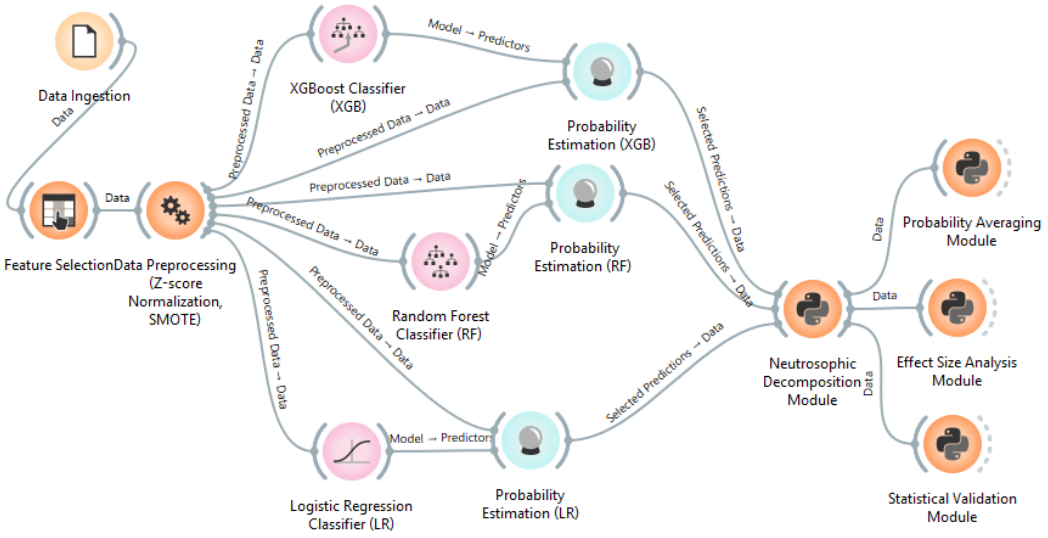

Figure 2. Orange Data Mining workflow implementing the proposed framework

As shown in Figure 2, the workflow routes preprocessed data through three parallel Predictions widgets—one per classifier—whose probability outputs converge in a Python Script node that computes the neutrosophic tripleta (T, I, F) and applies the adaptive abstention logic.

From the soft-voted distribution $P(\bar{x}) = (1/M)\sum P_m(x)$, we compute confidence magnitude (T/F) and uncertainty geometry (I):

$$T(x) = max_k \bar{P}_k(x) \quad (1)$$

$$F(x) = 1 - T(x) \quad (2)$$

$$I_{\text{entropy}} = H(\bar{P})/log(K), H(\bar{P}) = -\sum_k \bar{P}_k log\, \bar{P}_k \quad (3)$$

$$I_{margin} = 1 - (p_1 - p_2),\ p_1 \geq p_2 \text{ are top-2 probabilities} \quad (4)$$

$$I_{\text{gini}} = \left(1 - \sum_k \bar{P}_k^2\right)/(1 - 1/K) \quad (5)$$

Since F already encodes the complement of T, we empirically verify that I provides incremental information through four tests: (1) logistic regression comparing error ~ F versus error ~ F + I; (2) band-stratified correlation within fixed F ranges; (3) matched-pair comparison within F ∈ [0.3, 0.5]; (4) partial correlation r(I, error | F).

We evaluate on the IoT-CAD dataset [12], a digital forensics benchmark for cyberattack attribution in IoT environments (UNSW Canberra). After filtering rare classes, the dataset comprises 267,346 instances across eight categories: Normal, MITM, Injection, DoS, DDoS, Password, Malware, and Probing. Preprocessing includes StandardScaler normalization, SMOTE oversampling [13], and 80/20 stratified splitting. The test set comprises 82,332 instances. Three classifiers are trained: Random Forest, XGBoost (L1/L2 regularization), and Logistic Regression (L2, balanced weights).

4. Results and Analysis

The soft-voted ensemble achieves 97.0% overall accuracy (73,911 correct, 8,421 incorrect). Per-class F1-scores exceed 0.97 for most attack categories, with Normal traffic showing the lowest recall (0.91).

Table 1 presents discriminative power. F achieves the highest Pearson r (0.609) and Cohen's d (2.535) as a strong baseline. I_margin is closest (r = 0.601, d = 2.483), while I_entropy and I_gini trail slightly.

Table 1. Confidence magnitude (F) vs. Uncertainty geometry (I): discriminative power for error detection.

Metric	Mean (Correct)	Mean (Error)	Pearson r	Cohen's d
F (baseline)	0.1222	0.5667	0.6092	2.535
I_{entropy}	0.1621	0.5938	0.5452	2.146
I_{margin}	0.1916	0.8570	0.6012	2.483
I_{gini}	0.1979	0.7500	0.5565	2.210

The high correlation between I_{margin} and F (r = 0.990) raises a critical question: if they share 99% of linear variance, does I provide information that F does not? We address this through four analyses.

Logistic regression. Model A (error ~ F) achieves AIC = 27,335. Adding I_{margin} reduces AIC by 573 points ($p < 10^{-127}$). $I_{entropy}$ also contributes (ΔAIC = +311) but with a negative coefficient ($\beta = -1.16$), meaning higher entropy predicts fewer errors when F is controlled. I_{gini} does not contribute (p = 0.79).

Table 2. Logistic regression: incremental value of I beyond F.

Model	β_F	β_I	AIC	ΔAIC
error ~ F	2.160 ***	—	27,335	(baseline)
error ~ F + I_{margin}	0.801 ***	1.525 ***	26,761	+573
error ~ F + I_{entropy}	3.145 ***	−1.162 ***	27,023	+311
error ~ F + I_{gini}	2.180 ***	−0.026 n.s.	27,337	−2

Band-stratified analysis. Within fixed bands of F, I_{margin} maintains significant correlation with error. The effect is strongest in the critical decision zone F ∈ [0.3, 0.5], where r(I_{margin}, error) reaches 0.218–0.228—precisely where abstention decisions are most consequential.

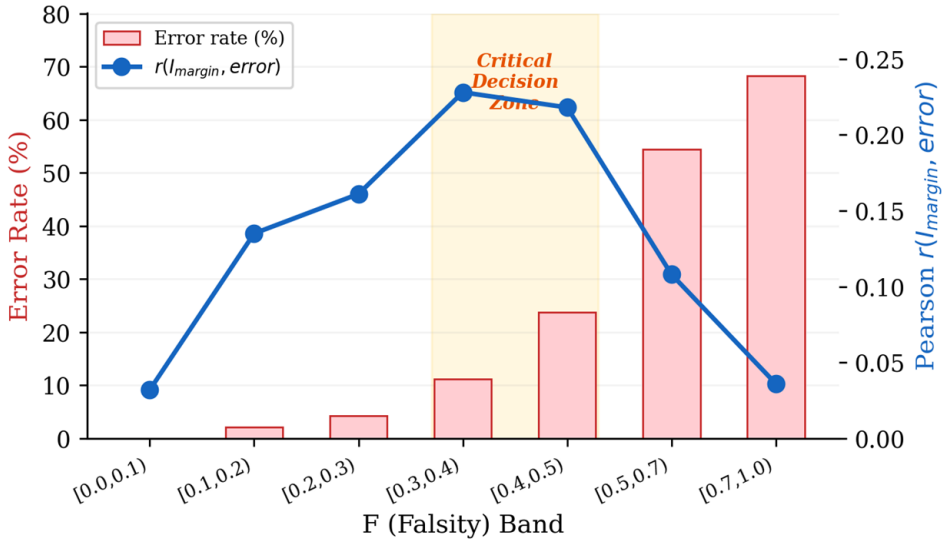

Figure 3. Error rate (bars) and I_{margin} correlation (line) within F bands.

Matched-pair comparison. Within F ∈ [0.3, 0.5] (N = 7,356), instances split by median I_{margin} show dramatically different error rates: 24.4% (high I_{margin}) versus 8.8% (low I_{margin}), a difference of 15.7 percentage points (Mann–Whitney $p < 10^{-114}$) despite near-identical mean F values (0.350 vs. 0.433).

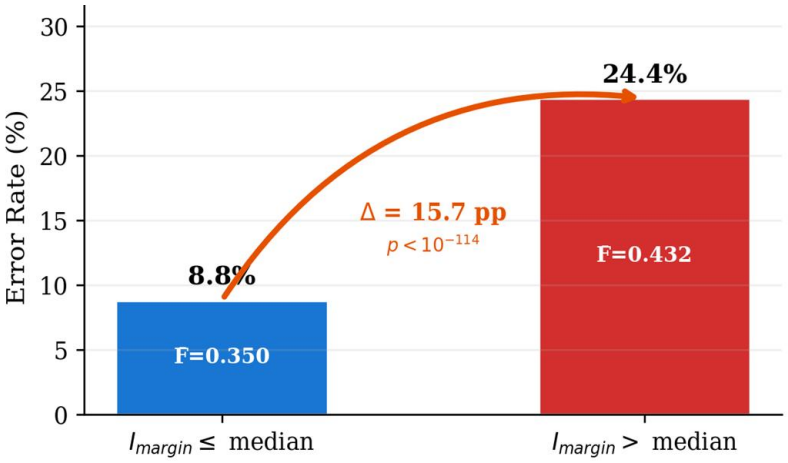

Figure 4. Matched-pair comparison within F ∈ [0.3, 0.5]: error rate by I_{margin} group.

This 15.7 percentage point gap constitutes the most direct evidence that I captures risk invisible to F. A final test examines whether this effect survives linear conditioning.

Partial correlation. $r(I_{margin}, \text{error} \mid F) = -0.019$ is statistically significant ($p < 10^{-8}$) but negligible in magnitude. This contradiction with the above tests is resolved by recognizing that partial correlation captures only linear effects, while I_{margin}'s contribution is fundamentally nonlinear—strongest in the intermediate F range and irrelevant at the extremes.

Conditional on F, I_{entropy} has a negative regression coefficient (β = −1.16). Two instances with F = 0.40 illustrate why: Scenario A has $\vec{P}$ = (0.60, 0.40, 0, ..., 0)—one strong competitor (I_{margin} = 0.80, I_{entropy} = 0.33). Scenario B has $\vec{P}$ = (0.60, 0.057, ..., 0.057)—probability spread uniformly (I_{margin} = 0.46, I_{entropy} = 0.52). Scenario A is more dangerous because the competitor has substantial support, yet entropy assigns a higher score to B. I_{margin} correctly identifies A as higher risk.

Decile analysis confirms a monotonic relationship between indeterminacy and error: zero errors in D1 (most confident) and over 57% in D10 (most uncertain).

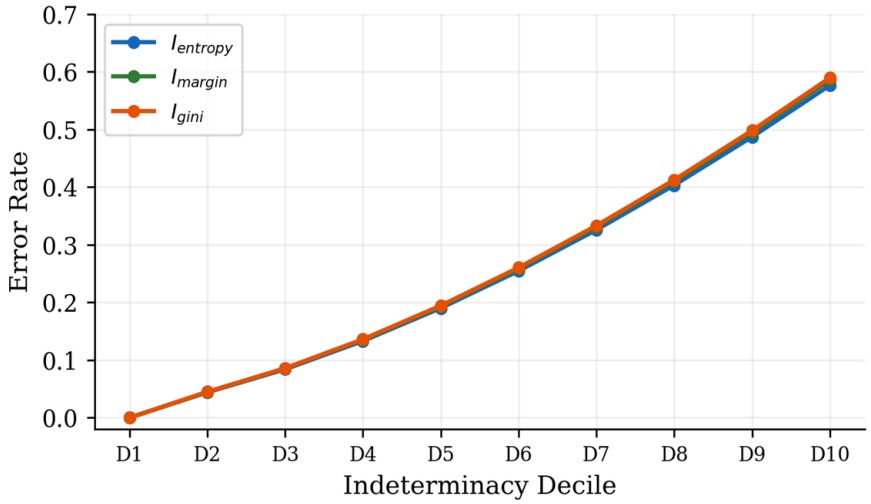

Figure 5. Error rate by indeterminacy decile (D1 = most confident, D10 = most uncertain)

A threshold sweep (Figure 6) demonstrates that selective abstention at $\tau = 0.4$ yields 99.2% accuracy at 75% coverage, with the Youden Index (Accuracy × Coverage) peaking at $\tau = 0.8$.

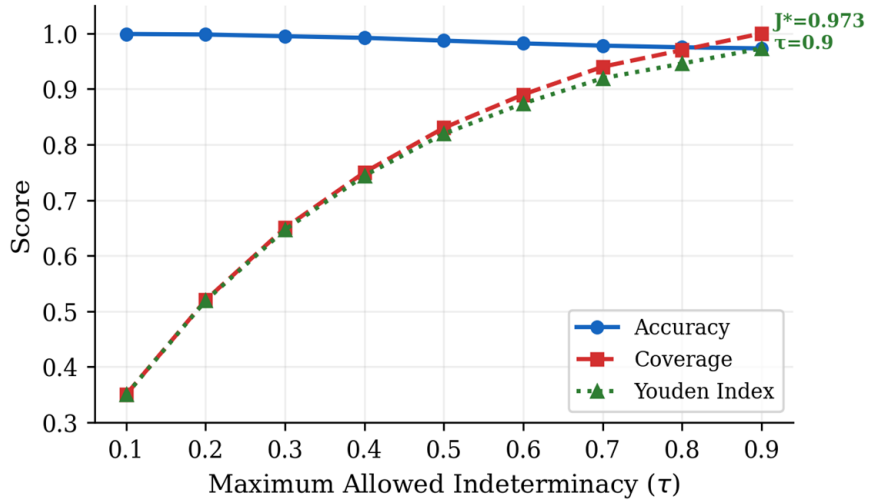

Figure 6. Accuracy, coverage, and Youden Index as a function of indeterminacy threshold

The tradeoff curves reveal three operational regimes: aggressive abstention ($\tau \leq 0.4$) achieves near-perfect accuracy (≥ 99.2%) at the cost of reviewing 25% of predictions; balanced operation ($\tau = 0.8$, Youden optimum) retains 97% of predictions with 97.5% accuracy; and full automation ($\tau = 0.9$) recovers baseline performance without abstention. The choice of τ depends on the deployment context: safety-critical IoT infrastructure favors low τ, while high-throughput monitoring environments favor the Youden optimum.

5. Discussion

Our central finding is that prediction reliability operates along two independent axes. Confidence magnitude (T/F) quantifies the overall confidence deficit. Uncertainty geometry (I) characterizes how residual mass is distributed among alternatives: concentrated on one strong competitor (dangerous) or diffused across many (benign). The practical implication is a two-stage decision protocol: (1) use T/F to identify predictions entering the uncertainty zone; (2) use I_{margin} to triage within that zone for human review.

In classical neutrosophic logic [4], T, I, and F are three independent quantities. In the classification context, however, T and F are deterministically linked (F = 1 − T), reducing

the tripleta to two independent axes: confidence magnitude (T ≡ F) and uncertainty geometry (I). T retains an operational role as the confidence score reported to the analyst. Future work could restore full three-dimensionality by deriving I from inter-model disagreement (e.g., Jensen–Shannon divergence) independently of the averaged distribution.

The equivalence $I_{\max}$ = F means not all neutrosophic formulations yield three independent quantities. For genuine three-dimensionality, I must capture distributional properties beyond the point maximum. Our results further refine this: $I_{g\text{ini}}$ adds no value beyond F (p = 0.79), while I_{margin} and I_{entropy} add significant but qualitatively different information. Researchers should verify both algebraic independence and empirical incrementality.

The analysis is conducted on a single dataset. SMOTE may introduce artifacts. The Orange implementation requires embedded Python. The two-stage protocol has not been evaluated in a live-edge deployment.

6. Conclusions

This paper addressed whether neutrosophic indeterminacy provides information beyond falsity for IoT intrusion detection. On the IoT-CAD dataset (82,332 instances, 97% accuracy), we demonstrated that confidence magnitude (T/F) is the strongest linear predictor of error (r = 0.609) but that uncertainty geometry (I_{margin}) contributes significant nonlinear incremental value (ΔAIC = 573), concentrated at the decision boundary where abstention matters most. Within this zone, high I_{margin} triples the error rate despite similar confidence levels (Δ = 15.7 pp, $p < 10^{-114}$). Confidence magnitude determines whether a prediction is uncertain; uncertainty geometry determines how dangerous that uncertainty is—justifying the neutrosophic tripleta precisely at the abstention boundary where it matters most for trust-aware IoT security.

Author Contributions

Conceptualization, M.L. and F.S.; methodology, M.L.; software, M.L.; validation, M.L., L.C. and A.M.P.; formal analysis, M.L.; investigation, M.L. and L.C.; writing—original draft preparation, M.L.; writing—review and editing, L.C., A.M.P. and F.S.; supervision, F.S. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The IoT-CAD dataset is publicly available on Kaggle (https://www.kaggle.com/datasets/hanya91/iot-cad-an-ai-based-iot-forensics-dataset) and IEEE DataPort (https://doi.org/10.21227/9age-4d85). The Orange Data Mining workflow and Python scripts for neutrosophic decomposition, statistical validation, and incremental analysis are archived at Zenodo (https://doi.org/10.5281/zenodo.19368604).

Acknowledgments

During the preparation of this manuscript, the authors used Claude (Anthropic, 2025) for the purposes of code generation for the Orange Data Mining pipeline scripts and document formatting. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest

The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript: IDS—Intrusion Detection System; IoT—Internet of Things; T—Truth; F—Falsity; I—Indeterminacy; RF—Random Forest; XGB—XGBoost; LR—Logistic Regression; SMOTE—Synthetic Minority Over-sampling Technique; AIC—Akaike Information Criterion.

References

1. Khraisat, A.; Gondal, I.; Vamplew, P.; Kamruzzaman, J. Survey of intrusion detection systems: Techniques, datasets and challenges. Cybersecurity 2019, 2, 20. https://doi.org/10.1186/s42400-019-0038-7.
2. Sommer, R.; Paxson, V. Outside the closed world: On using machine learning for network intrusion detection. In Proceedings of the 2010 IEEE Symposium on Security and Privacy, Oakland, CA, USA, 16–19 May 2010; pp. 305–316. https://doi.org/10.1109/SP.2010.25.
3. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. arXiv 2016, arXiv:1606.06565.
4. Smarandache, F. A unifying field in Logics: Neutrosophic Logic. In Philosophy; American Research Press: Rehoboth, NM, USA, 1999; pp. 1–141.
5. Al-Masri, E. Deciding When Not to Decide: Indeterminacy-Aware Intrusion Detection with NeutroSENSE. In Proceedings of the 2025 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 7–10 May 2025; pp. 0142–0148. https://doi.org/10.1109/AIIoT63789.2025.10841741.
6. Kavitha, B.; Karthikeyan, S.; Maybell, P.S. An ensemble design of intrusion detection system for handling uncertainty using Neutrosophic Logic Classifier. Knowl.-Based Syst. 2012, 28, 88–96. https://doi.org/10.1016/j.knosys.2011.12.004.
7. Kraipeerapun, P.; Fung, C.C.; Wong, K.W. Ensemble neural networks using interval neutrosophic sets and bagging. In Proceedings of the Third International Conference on Natural Computation (ICNC 2007), Haikou, China, 24–27 August 2007; Volume 1, pp. 386–390. https://doi.org/10.1109/ICNC.2007.511.
8. Akbulut, Y.; Şengür, A.; Guo, Y.; Smarandache, F. A novel neutrosophic weighted extreme learning machine for imbalanced data set. Symmetry 2017, 9, 142. https://doi.org/10.3390/sym9080142.
9. Dietterich, T.G. Ensemble methods in machine learning. In Multiple Classifier Systems; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1.
10. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems; Curran Associates: Red Hook, NY, USA, 2017; Volume 30.
11. Demšar, J.; Curk, T.; Erjavec, A.; Gorup, Č.; Hočevar, T.; Milutinovič, M.; Možina, M.; Polajnar, M.; Toplak, M.; Starič, A.; et al. Orange: Data Mining Toolbox in Python. J. Mach. Learn. Res. 2013, 14, 2349–2353.
12. Mohamed, H.; Koroniotis, N.; Schiliro, F.; Moustafa, N. IoT-CAD: A comprehensive Digital Forensics dataset for AI-based Cyberattack Attribution Detection methods in IoT environments. Ad Hoc Netw. 2025, 174, 103840. https://doi.org/10.1016/j.adhoc.2025.103840.
13. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 2002, 16, 321–357. https://doi.org/10.1613/jair.953.
14. Cohen, J. Statistical Power Analysis for the Behavioral Sciences, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1988.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.